Research Report 1430

Preliminary Report on a National Cross-Validation of the Computerized Adaptive Screening Test (CAST)

Deirdre J. Knapp and Rebecca M. Pliske

Personnel Utilization Technical Area

Manpower and Personnel Research Laboratory

U. S. Army

Research Institute for the Behavioral and Social Sciences

August 1986

Approved for public release; distribution unlimited.

BLANK PAGES
IN THIS
DOCUMENT
WERE NOT
FILMED

U. S. ARMY RESEARCH INSTITUTE

FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

WM. DARRYL HENDERSON
COL, IN
Commanding

Technical review by

Clessen J. Martin
Elizabeth P. Smith

NOTICES

DISTRIBUTION. Primary distribution of this report has been made by ARI Please address correspondence concerning distribution of reports to U.S Army Research Institute for the Behavioral and Social Sciences, ATTN PERI-POT, 5001 Eisenhower Ave Alexandria, Virginia 22333-5600

FINAL DISPOSITION This report may be destroyed when it is no longer needed Please do not return it to the U S Army Research Institute for the Behavioral and Social Sciences

NOTE The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARI Research Report 1430	2. GOVT ACCESSION NO. AD-A175 767	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PRELIMINARY REPORT ON A NATIONAL CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)		5. TYPE OF REPORT & PERIOD COVERED Interim Report January 1985-March 1986
		6. PERFORMING ORG. REPORT NUMBER --
7. AUTHOR(s) Deirdre J. Knapp and Rebecca M. Pliske		8. CONTRACT OR GRANT NUMBER(s) --
9. PERFORMING ORGANIZATION NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Ave., Alexandria, VA 22333-5600		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q263731A792 221H3
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Ave., Alexandria, VA 22333-5600		12. REPORT DATE August 1986
		13. NUMBER OF PAGES 33
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) --		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Recruiting
Computerized Adaptive Testing (CAT)
Computerized Adaptive Screening Test (CAST)
Enlistment Screening Test (EST)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The Computerized Adaptive Screening Test (CAST) is used by Army recruiters to predict prospective applicants' (i.e., prospects') subsequent performance on the Armed Forces Qualification Test (AFQT). A modified version of the CAST software was used in 60 recruiting stations across the country from January through December 1985 to collect cross-validation data. Recruiters in these test stations also recorded test scores for prospects given CAST's paper-and-pencil counterpart, the Enlistment Screening Test (EST). All test data were matched to applicant tapes from Military Entrance Processing

(Continued)

ARI Research Report 1430

20. (Continued)

Stations (MEPS) to obtain AFQT scores and relevant demographic data. Analyses indicate that both CAST and EST are good predictors of AFQT performance. Alternative CAST subtest lengths were also examined, and the current operational subtest lengths ·ppear to be optimal. Analyses of CAST's accuracy in predicting prospects' subsequent classification into important AFQT categories (i.e., 1-3A and 3B) are also discussed.

Research Report 1430

Preliminary Report on a National Cross-Validation of the Computerized Adaptive Screening Test (CAST)

Deirdre J. Knapp and Rebecca M. Pliske

Personnel Utilization Technical Area
Paul A. Gade, Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

August 1986

Army Project Number
2Q263731A792

Manpower and Personnel

FOREWORD

The Army must meet quantity and quality goals in its recruiting. Recent advances in computer technology and psychometric theory have made possible a new type of assessment technique, called computerized adaptive testing (CAT), that can provide accurate estimates of ability based on relatively few test items. The Computerized Adaptive Screening Test (CAST) was designed to estimate a prospect's Armed Forces Qualification Test (AFQT) score at the recruiting station. Recruiters use prospects' CAST scores to determine whether applicants should be sent to Military Entrance Processing Stations for further testing. These scores also forecast the various options and benefits for which the prospects will subsequently qualify. This report summarizes analyses from a nation-wide cross-validation study and recommends that changes be made to CAST to improve its utility to recruiters. Keywords: aptitude tests; test construction (psychology)

Edgar M Johnson
EDGAR M. JOHNSON
Technical Director

PRELIMINARY REPORT ON A NATIONAL CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

EXECUTIVE SUMMARY

Requirement:

To cross-validate the Computerized Adaptive Screening Test (CAST) with a nationally representative sample of prospective applicants (prospects) and to provide recommendations on how to improve the utility of CAST to field recruiters.

Procedure:

A modified version of the CAST software was used in 60 recruiting stations across the country from January through December 1985 so that prospects' CAST performance could be recorded on data diskettes for analysis. Recruiters recorded test scores and social security numbers (SSNs) for those prospects given the Enlistment Screening Test (EST) instead of CAST, and these data were also forwarded to ARI. Both CAST and EST scores were matched by SSNs to applicant tapes from Military Entrance Processing Stations to obtain AFQT scores and relevant demographic data. These data were examined, using regression and cross-tabulation analyses.

Findings:

The findings presented in this report are based on the first 6 months of data collection and indicate that the current operational version of CAST is quite good at predicting AFQT scores (r=.82). The efficiency of the current length of the test, 10 Word Knowledge (WK) and 5 Arithmetic Reasoning (AR) items, surpasses that of other alternatives (e.g., 10 WK and 10 AR items). Race and sex subgroup differences in AFQT predictions based on CAST exist, but the magnitude of these differences is not large. The analyses of the EST data indicate that it is also a very good predictor of AFQT performance (r=.79). Analyses of CAST's accuracy at predicting prospects' subsequent classification into important AFQT categories (e.g., categories 1-3A) indicate that, although the current version of CAST does a good job at category prediction, this type of prediction could be improved. To this end, the report recommends modifying the CAST software to provide recruiters with probabilistic information about prospects' subsequent classification into AFQT categories.

Utilization of Findings:

This report will be used by the U.S. Army Recruiting Command to make decisions about future modifications to the CAST software.

PRELIMINARY REPORT ON A NATIONAL CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

PRELIMINARY REPORT ON A NATIONAL CROSS-VALIDATION OF THE COMPUTERIZED ADAPTIVE SCREENING TEST (CAST)

INTRODUCTION

The Computerized Adaptive Screening Test (CAST) was designed by the Navy Personnel Research and Development Center (NPRDC) under the sponsorship of the Army Research Institute (ARI) to provide a prediction of prospects' Armed Forces Qualification Test (AFQT) scores at recruiting stations. The purpose of this report is to review CAST's background and to describe a large scale CAST data collection effort. A discussion of statistical analyses that have been performed on the first six months of the twelve month data collection will provide insight into how CAST is performing and how it might be changed to optimize its usefulness.

Background

Individuals interested in joining any of the armed services are required to take the Armed Services Vocational Aptitude Battery (ASVAB). ASVAB scores are used to determine eligibility for enlistment and to assist in determining initial training assignments. The ASVAB is administered under secure testing conditions either by the Department of Defense High School Testing Program or at a Military Entrance Processing Station (MEPS) or Mobile Examining Team (MET) site. Most testing is conducted at MEP/MET locations. Sending individuals to these sites represents a significant financial investment for the armed services. In addition to the costs of the testing itself; travel, lodging, and boarding expenses are typically incurred. Both the recruiter and the prospect also invest a significant amount of time in this process. The recruiter must make arrangements to insure that the prospect gets to the testing site. For the prospect, the three and one-half hours required to take the test battery must be added to the time spent getting to and from the testing location.

AFQT scores are currently computed by adding together four ASVAB subtest scores. Specifically, word knowledge (WK), arithmetic reasoning (AR), paragraph comprehension (PC) subtest scores, and one-half of the numerical operations (NO) subtest score combine to produce AFQT. An individual's AFQT score is intended to reflect his or her "trainability." Thus, AFQT scores are used to assess the eligibility of applicants for enlistment and special benefits. Applicants for the Army who score at or above the 50th percentile (AFQT categories 1, 2, and 3A) are eligible for special options and benefits such as the 2-year Enlistment Option and the Army College Fund. Applicants who score between the 31st and 49th percentiles on AFQT (AFQT category 3B) qualify for enlistment but are not eligible for special options. Those individuals who score between the 16th and 30th percentiles (AFQT categories 4A and 4B) are generally regarded as being low priority recruits. In fact, the Army is currently not accepting individuals who score below the 26th percentile.

Thus there are two major reasons why information that predicts prospects' AFQT performance is invaluable to Army recruiters. A test that provides this kind of information can be used simultaneously as an informal screening device and a sales tool. If the test indicates that a prospect has very little chance of subsequently qualifying for enlistment, the recruiter may choose to discourage him or her from further interest in the Army. Besides saving the expense of ASVAB testing, this allows recruiters to spend a greater amount of time selling the Army to more promising prospects. One of the major functions of a recruiter is to convince qualified prospects that the Army is a desirable job alternative. The special options and benefits offered by the Army are powerful incentives, but they only work if an individual subsequently qualifies for them. In other words, their utility depends upon the recruiter using them with the right people. Clearly, a test that predicts subsequent AFQT performance gives recruiters the information they need to most effectively perform their jobs.

The Enlistment Screening Test (EST) is currently available to all of the armed services for use at recruiting stations as a predictor of AFQT performance. Although EST provides fairly accurate predictions of AFQT scores, it has several drawbacks that it shares with most other paper-and-pencil tests. The major drawbacks concern administration time, clerical burden, and scoring errors (cf. Baker, Rafacz, & Sands, 1984). Recruiters must allow prospects 45 minutes to complete EST and then they must hand score the test. This latter step takes additional time and is subject to error. Because there are only two alternative EST forms, it is possible that prospective applicants might learn the items and eventually pass the test on repeated testing in different recruiting stations. Excessive test time, clerical burden, and test security are problems that can be alleviated or eliminated because of recent advances in computer technology and psychometric theory.

Computerized Adaptive Testing

An advance in psychometric theory, called Item Response Theory (IRT), has made it possible to adapt or "tailor" a test to the individual examinee (Lord, 1980). Unlike ability tests based on classical test theory, ability tests based on IRT can provide comparable estimates of individuals' ability levels even when different individuals receive different sets of test items. In classical test theory all test parameters, such as item difficulty and discrimination indices, are dependent on the specific test (i.e., a specific combination of items) and on the characteristics of the sample of individuals with whom the test was developed. In IRT, the focus is on individual test items and the probability of correct response to each item given a specific ability level. The estimate of an individual's ability level is based on parameters associated with the specific items that individual received; these parameters are independent of the other items on the test and are also independent of the characteristics of the developmental sample. A detailed discussion of IRT is beyond the scope of this report. The interested reader is referred to Warm (1978) for an excellent introduction to IRT.

In traditional tests, each examinee responds to all items on the test. The traditional approach to test construction results in relatively poor measurement at the high and low ability extremes because many items on the test tend to be too difficult for the low ability examinees or too easy for the high ability examinees. In adaptive testing, each examinee receives the items that are appropriate to his or her ability level. The selection of each subsequent item is based on the examinee's previous response. If an examinee has responded correctly to the previous item, then the next item will usually be more difficult than the previous one. If the examinee's response to the previous item was incorrect, then the next item will usually be easier than the previous one. Adaptive testing makes it possible to construct tests that are able to discriminate equally well across all ability levels.

Although adaptive testing is possible without a computer, it is not very practical because of the number of calculations and branching decisions that need to be made. In computerized adaptive testing, the computer presents each item and records the examinee's response. It computes an estimate of the examinee's ability level that determines the item that is administered next. A detailed discussion of some of the alternative procedures for making ability estimates and selecting subsequent items can be found in a report by McBride (1979).

In addition to improving the discriminability of a test, computerized adaptive tests are more efficient to use than traditional paper-and-pencil tests because they reduce testing time without sacrificing validity. Computerized adaptive tests also eliminate the need for manual scoring and recording which can result in clerical errors, and they can provide immediate feedback on test results. Computerized adaptive tests reduce test compromise by eliminating test booklets which can be stolen, and by administering different items to different individuals making it more difficult for individuals to "cheat." For all of these reasons, a computerized adaptive test that can accurately predict a prospect's AFQT score is a highly desirable recruiting tool; thus, the Computerized Adaptive Screening Test (CAST) was developed.

Development of CAST

The item pools for CAST were constructed by researchers at the University of Minnesota (cf. Moreno, Wetzel, McBride, & Weiss, 1983) for use in the initial developmental stages of a computerized adaptive version of ASVAB (called CAT ASVAB). Moreno, et al. provided a de facto pilot test of CAST in their research which examined the relationship between corresponding ASVAB and CAT ASVAB subtests. These researchers administered the WK, AR, and PC subtests to 270 male Marine recruits at the Marine Corps Recruit Depot in San Diego, California. The data from this pilot test yielded a correlation of .87 between the three optimally-weighted CAT ASVAB subtests and ASVAB AFQT. Because the statistical analyses indicated that the PC subtest did not contribute a significant amount of predictive power beyond that provided by the WK and AR subtests, and because the PC subtest items required an inordinate amount of time to administer, this subtest was dropped from CAST.

Presently, there are 78 items in CAST's WK item pool and 225 items in CAST's AR item pool. All items are multiple choice with a maximum of five response alternatives. CAST uses a three-parameter logistic ogive item response model (Birmbaum, 1968); thus each item has three parameters (discrimination, difficulty, and guessing) associated with it. Test items for CAST item pools were chosen so that the discrimination parameter values would be greater than or equal to .78; the difficulty parameter values would range between +2 and -2; and the guessing parameter values would be less than or equal to .26. CAST uses the Bayesian sequential scoring procedure discussed by Jensema (1977) to score and select subsequent items for administration. The test ends when the examinee has responded to 10 WK and 5 AR items.

Prior Validation Efforts

Typically, the validity of a test like CAST is estimated by computing the correlation between corresponding sets of predictor (e.g., CAST) and criterion (e.g., AFQT) scores. A correlation coefficient reflects the amount of variability in one set of scores that can be accounted for, or explained by, another set of scores. Its value can range from +1.0 to -1.0, with a value of zero indicating that there is no correspondence between the two sets of scores. In the following discussion, reference will be made to bivariate correlations and to multiple correlations. A bivariate correlation (*r*) is computed when only one "predictor" is used (e.g., total CAST score). When multiple predictors (e.g., WK and AR subtest scores) are used, then a multiple correlation (*R*) is the relevant statistic to report.

There are three validation efforts associated with CAST. The initial validation project was conducted at the Los Angeles MEPS with a sample of 312 U.S. Army applicants (Sands & Gade, 1983). Each applicant received 20 WK items and 15 AR items on an APPLE-II microcomputer. The data were analyzed to determine the optimal combination of subtest lengths so that the predictive accuracy of CAST would be at least as high as that estimated of EST (*r*=.83; Mathews & Ree, 1982) with the shortest administration time possible. Multiple correlation coefficients were computed for each of the 300 combinations of subtest lengths. Examination of the results led to the recommendation that the operational version of CAST be terminated following the administration of 10 WK and 5 AR items. The multiple correlation between this optimally-weighted subtest score combination and actual AFQT score was .85.

There were three major limitations to the initial validation of CAST. First, the statistical analyses were based on a relatively small sample. Second, the sample was not nationally-representative. Third, the testing environment was different from that in which CAST is actually used. CAST is administered on an individual basis in recruiting stations. In the initial validation project, CAST was administered by researchers to groups of examinees at a MEPS. Even if the sample and testing situation had been more appropriate, however, the need would still remain for at least one additional validation effort. Although it is always wise to derive estimates of a test's validity on more than one sample, there are circumstances where this extra step is mandatory. For example, in the present situation Sands and Gade (1983) conducted a regression analysis that produced weights for combining CAST's WK and AR subtests to optimize AFQT prediction. Because those weights capitalized on the chance variation in the data from which they were derived, the validity of the test scores based on those weights is overestimated. Therefore, it is necessary to test the predictive power of those weights on a second, independent sample. This step is called "cross-validation."

Army recruiting stations in the midwestern region of the U.S. provided CAST cross-validation data during January and February, 1984 (Pliske, Gade, & Johnson, 1984). CAST was introduced by geographical region, and the midwestern region was the only fully operational region at the time of data collection. Recruiters in these stations recorded prospects' CAST scores and social security numbers (SSNs) on log sheets. The U.S. Army Recruiting Command (USAREC) collected these data and forwarded them to ARI for analysis. The CAST scores recorded by the recruiters were matched by SSNs to applicant tapes from the MEPS to obtain AFQT scores and relevant demographic data. Matching records were located for 1,962 individuals. The bivariate correlation coefficient between CAST and AFQT scores computed from these data was .80. This value reflects a reasonable amount of "shrinkage" from the original validity estimate of .85.

Although the validity estimates yielded by these two projects suggest that CAST is an effective predictor of AFQT, an additional data collection effort was required. Specifically, a large scale cross-validation effort using a representative sample of all Army prospects was called for. Such an effort was undertaken in January 1985 and was completed in December 1985. The data collected in this project will also suggest ways that CAST could be changed to optimize its usefulness to Army recruiters.

PROCEDURE

Data Collection Procedure

A modified version of the CAST software was designed for use in this latest validation project. The program was changed so that examinees' test responses would be recorded on special data collection floppy diskettes. Information recorded on the data diskettes included item identification number, examinees' answer, the time it took for the examinee to read and answer the item, and the examinee's SSN. Each diskette recorded test information for many examinees. The software was also changed so that the prospects would respond to 15 WK and 10 AR items. However, the predicted AFQT score reported at the end of the test was based on the operationally-used stopping rule of 10 WK and 5 AR items.

CAST data diskettes were collected from 60 recruiting stations located across the country. These stations were selected to be representative of the population of approximately 2,000 Army recruiting stations in terms of geographic location and population density. A full year of data collection was required to insure that the sample of prospects would not be biased by seasonal fluctuations in prospect characteristics.[1] The analyses discussed in this paper are based on data collected during the first six months of this project. Thus all results should be considered preliminary pending verification with the entire 12 months of data.

Army recruiters use EST rather than CAST when they do not have access to their JOIN microcomputer systems. Because EST has never been cross-validated, this was an ideal opportunity to do so. Consequently, the 60 participating recruiting stations were also given log sheets to record the predicted AFQT scores and SSNs of prospects to whom they administered EST. The EST log sheets were forwarded to ARI along with the CAST data diskettes at the end of each month.

When the CAST diskettes were received at ARI, the data from the individual recruiting stations were concatenated into a larger data set and uploaded to an IBM mainframe computer system where the data could be analyzed. The EST data were directly keyed into the mainframe system. Once on the mainframe, the information recorded at recruiting stations was matched to information available on MEPS applicant tapes.

Before describing the samples, problems encountered in the data collection effort will be noted. Although the majority of recruiting stations selected to participate in this project consistently submitted CAST data diskettes at the end of each month, the participation of several stations was sporadic. The submission of EST log sheets was less consistent than the submission of CAST data diskettes. Many of the scores received by ARI were not matched to MEPS records because incorrect SSNs were recorded at the recruiting stations. Also, many cases were lost because a large number of prospects never went to MEPS for further testing. Table 1 illustrates the severity of these problems with regard to the CAST data.

Table 1

Percentage of CAST Scores Successfully Matched to AFQT Scores

by Source of CAST Scores

Brigade	Number of CAST Scores*	Percentage of CAST Scores Matched to AFQT
First	827	24%
Second	2014	33%
Fourth	753	40%
Fifth	1678	40%
Sixth	1053	34%

*Total number of CAST cases was 6470. Brigade was not identifiable for 2% of the cases.

[1] The analyses of seasonal differences in prospect characteristics will be presented in the final report for this data collection effort.

Sample Characteristics

Tables 2 and 3 summarize the major demographic characteristics of the CAST and EST samples. The most significant difference between the two samples is their size. It is difficult to determine the extent to which these samples accurately represent the population of Army prospects because no data are available to accurately describe that population. It is likely, however, that the samples exhibit differences from the Army prospect population because many prospects fail to go to MEPS for ASVAB testing and our samples are based only on those prospects for whom we located a matching MEPS record. On the basis of the information provided to them by recruiters, some prospects decide that they are not interested in joining the Army so they do not go to MEPS. Further, recruiters choose not to encourage some prospects to go to MEPS because their prequalification information suggests that the prospects are unsuitable for enlistment in the Army. Thus certain kinds of prospects are being systematically excluded from the CAST and EST samples. Note that the systematic exclusion of lower ability and higher ability prospects leads to a restriction in range of AFQT and CAST/EST scores. This, in turn, results in estimates of correlation that will be somewhat lower than appropriate.

Given the absence of more appropriate criteria, the adequacy of these samples can be evaluated in terms of the sample selection procedure and common sense expectations. The experimental recruiting stations were selected to be representative of all recruiting stations in terms of geographical location and population density. Because blacks represent a small percentage of the population of American citizens, the sample selection procedure was also designed to insure that a relatively large number of black prospects would be included in the samples. A sufficient number of black prospects was needed to permit legitimate comparison to white prospects. Other characteristics (e.g., average age and percentage of males) of the samples correspond quite well with a priori expectations.

RESULTS AND DISCUSSION

CAST Validation Information

The CAST validity estimates from the present investigation and the two validation projects described earlier are shown in Table 4. A correlation of .82 indicates that there is a strong, linear relationship between CAST and AFQT scores. The corresponding coefficient of determination (r^2=.67) indicates that we can account for approximately 67% of the variability in applicants' AFQT scores by knowing their CAST scores; however, this coefficient also shows that approximately 33% of the variance in applicants' AFQT scores is left unaccounted for by knowledge of performance on CAST. It is important to note that the validity of CAST could never exceed the test-retest reliability of ASVAB AFQT which is about .90. Thus even if applicants were given the entire ASVAB to predict their performance on a subsequent administration of ASVAB, approximately 19% of the variability in their scores on the second administration would be unaccounted for by knowledge of their score on the first administration.

Table 2

CAST Sample Description

Sample Size	2,240
Sex	81% Male
	19% Female
Race	59% White
	37% Black
	4% Other
Age	Mean=20; SD=3.47
	Median=19
	Mode=18
Component	85% Regular Army
	15% Army Reserve
Education	4% Some College/Vo Tech
(Based on 65% Cases)	77% HS Diploma or GED
	19% Non-HS Graduates*
AFQT Category	23% 1 and 2
(From ASVAB)	15% 3A
	28% 3B
	34% 4A-5

*Includes high school seniors

Table 3

EST Sample Description

Sample Size	688
Sex	84% Male
	16% Female
Race	62% White
	34% Black
	4% Other
Age	Mean=20; SD=4.14
	Median=19
	Mode=18
Component	78% Regular Army
	22% Army Reserve
Education	5% Some College/Vo Tech
(Based on 73% Cases)	71% HS Diploma or GED
	23% Non-HS Graduate*
AFQT Category	24% 1 and 2
(From ASVAB)	15% 3A
	29% 3B
	32% 4A-5

*Includes high school seniors

Table 4

Correlation Between Current Operational Version of CAST and AFQT

Sample	N	r	r^2
LA MEPS (1983)	312	.85	.72
4th Brigade (1984)	1,962	.80	.64
National (1985)	2,228	.82	.67

Factors that might explain variance in AFQT scores beyond that accounted for by CAST include test anxiety, noisy test environments, and fatigue. These influences on test performance introduce random error into AFQT and CAST scores that defies identification and control. All tests are characterized by this type of random error, so no test can work as a perfect predictor. However, some of the AFQT variance left unexplained by CAST scores might be attributable to systematic influences of factors such as educational background or ethnic group membership. This possibility was investigated by computing a series of multiple correlation coefficients in which variables were added into the regression equation one by one. Table 5 summarizes the results of this analysis. The only variables that add to the predictive power of CAST are sex and race (either Black or White)[2].

Table 5

Percent of Variance Accounted for by Stepwise Addition

of Variables to Regression Model (CAST)

Predictor	R^2
CAST Score	.667
Race	.676
Sex	.681
Years of Education	.682
Age	.684
ASVAB Version	.685

[2] Data from other ethnic groups were not included in any analyses dealing with race because members of these groups are not adequately represented in either the CAST or EST samples.

Although the increase in explained variance due to race and sex is small, the influence of these variables is examined in more detail in a report by Knapp, Pliske, & Elig (1986). The report concludes that there are differences between race and sex subgroups in the way in which CAST performs. Generally, the AFQT performance of black prospects tends to be overpredicted by CAST relative to white AFQT performance predictions. This means there is a slight tendency for the CAST scores of black prospects to predict they will perform better on AFQT than they actually do. The AFQT performance of female prospects tends to be underpredicted by CAST relative to male AFQT performance predictions, which means that there is a slight tendency for the CAST scores of female prospects to predict they will perform more poorly on AFQT than they actually do. These differences for the racial and sexual subgroups, however, are not large and should not be cause for undue concern by Army policy makers.

Subtest Length

The data collected in this project provide the information required to reevaluate the number of subtest items administered by CAST. The initial decision of 10 WK and 5 AR items was based on validity estimates that came from a relatively small sample (Sands & Gade, 1983). In the present project, validity estimates based on a larger, more representative sample can be computed. Further, the average time it takes to administer various subtest length combinations can be determined.

Table 6 lists the validity estimates and administration times associated with selected subtest length combinations. Examination of this table shows that rather substantial increases in administration time are required to significantly increase the predictive accuracy of the test. For example, it takes an average of slightly over 12 minutes to administer the current operational version of CAST that has a predictive accuracy (or $\underline{R}^2$ value) of .67, and it would require a 6-minute increase in administration time to improve the predictive accuracy by 5% (i.e., $\underline{R}^2$=.72). These results corroborate earlier evidence suggesting that optimal efficiency is achieved with 10 WK and 5 AR items.

Table 6

Multiple Correlations Between AFQT and Selected CAST Subtest Lengths

Number WK	Number AR	R	R^2	Average Administration Time (in Minutes)*
5	5	.79	.62	10.10
** 10	5	.82	.67	12.15
15	5	.83	.69	14.24
5	10	.82	.67	14.01
10	10	.84	.70	16.07
15	10	.85	.72	18.16

* Administration time includes time spent reading instructions and taking practice items.
** Operational test length.

EST Validation Information

The original validation of EST resulted in a validity estimate of .83 (Mathews & Ree, 1982). This estimate was derived from a sample of 486 prospects who completed test booklets in either Army, Navy, Air Force, or Marine recruiting stations. Based on data collected in the first six months of the present project, the correlation between EST and AFQT scores is .79. Judging from this latter estimate, EST is explaining approximately 62% of the variability in AFQT performance and 38% of the variance is not accounted for by knowledge of EST performance. A stepwise regression analysis similar to the one presented in Table 5 for the CAST data was performed on the EST data in an attempt to identify factors that explain AFQT variance beyond that accounted for by EST performance. The results of this analysis appear ir Table 7. Race and years of education add a small amount of predictive information, but the other variables seem to have virtually no impact on the ability to predict AFQT performance.

Table 7
Percent of Variance Accounted for by Stepwise Addition of Variables to Regression Model (EST)

Predictor	Explained Variance (R^2)
EST Score	.629
Race	.638
Years of Education	.650
Age	.652
Sex	.654
EST Version (81A/81B)	.655
ASVAB Version	.655

Even though the information in Table 7 indicates that examinee sex is not a factor that influences the performance of EST, in the interest of test fairness, the impact of both sex and race are examined in detail in Knapp, et al. (1986). The analyses reported therein corroborate the evidence presented in Table 7. That is, examinee sex does not appear to influence the nature of EST predictions. As with CAST, however, the AFQT performance of black prospects tends to be overpredicted by EST relative to the AFQT performance predictions for white prospects. It should be noted that EST tends to indicate that all examinees will perform better on AFQT than they subsequently do. This tendency, however, is somewhat stronger for black examinees.

Category Prediction

At the present time, after the prospect completes CAST the computer presents bar charts that represent examinee performance on the WK and AR subtests and the examinee's predicted AFQT percentile score. There are several problems with this information. First, the predicted AFQT score currently provided by CAST is based on outdated norms and must be updated to predict ASVAB AFQT scores based on the 1980 Youth Norms. Second, additional information should be provided to help recruiters interpret CAST scores. Because the great majority of recruiters have never been taught the fundamentals of regression analysis, they do not adequately understand the nature of point predictions. Hence, recruiters expect predicted and actual AFQT scores to be exactly the same. As noted earlier, predictions based on statistical probabilities will always be somewhat imperfect. The extent to which a point prediction comes close to the actual value is reflected in the size of the standard error of estimate. The standard error of estimate for CAST is 13.6.

Recruiters use CAST to answer two questions. First, is the prospect likely to qualify as an adequate Army enlistee (AFQT category 3B or above)? Second, is the prospect likely to qualify for special options and benefits (AFQT category 3A and above)? Given this situation, it is not necessary to provide recruiters with AFQT score predictions. Rather, CAST could be modified to provide the two probabilities that answer the aforementioned questions (i.e., the probability that the examinee will subsequently be classified into category 3B or above on AFQT and the probability that the examinee will subsequently be classified into category 3A or above on AFQT).[3] The use of category predictions such as these would provide a recruiter with better information on which to base his or her decision whether to encourage or discourage the prospect from going on for further testing.

Currently, most recruiters probably make AFQT category predictions by interpreting CAST scores at face value. For example, a prospect who receives a predicted AFQT score of 31 would be predicted to belong to AFQT category 3B and a prospect with a predicted AFQT score of 30 would be predicted to belong to AFQT category 4A. Assuming that this is the way in which recruiters use CAST scores, this prediction scenario can be modeled statistically to determine how well CAST is currently predicting AFQT category classification.

Figure 1a shows CAST prediction results at the 3B/4A cutpoint when the assumption described above is made. This figure divides the sample of examinees into four groups: (1) those examinees correctly predicted by CAST to be members of ASVAB AFQT category 1-3B rather than category 4A-5, (2) those examinees incorrectly predicted by CAST to be members of ASVAB AFQT category 1-3B, (3) those examinees correctly predicted by CAST to be members of ASVAB AFQT category 4A-5 rather than category 1-3B, and (4) those examinees incorrectly predicted by CAST to be members of ASVAB AFQT category 4A-5. Out of the entire CAST sample, 79% of the examinees are correctly classified into either the 1-3B category or the 4A-5 category. The performance of most of the examinees misclassified by CAST was overpredicted. That is, when CAST was wrong, it was most likely to misclassify an unqualified examinee into the "passing" category.

In Figure 1b, the CAST prediction results at the 3A/3B cutpoint are shown. The predictions at this cutpoint are somewhat more accurate than those at the 3B/4A cutpoint (88% correct predictions versus 79% correct predictions). Furthermore, neither type of prediction error is predominate. In other words, at this cutpoint, CAST is just as likely to overpredict AFQT performance as it is to underpredict AFQT performance.

----

[3] ARI has provided USAREC with AFQT category prediction tables, based on earlier validation efforts, that state the probabilities of subsequent ASVAB AFQT category classification associated with the predicted AFQT score provided in the CAST output. These tables were to be distributed to recruiters to aid them in the interpretation of CAST scores. We are recommending that this type of information be directly incorporated into the CAST software.

PATTERN OF CAST PREDICTIONS AT TWO AFQT CATEGORY CUTPOINTS*

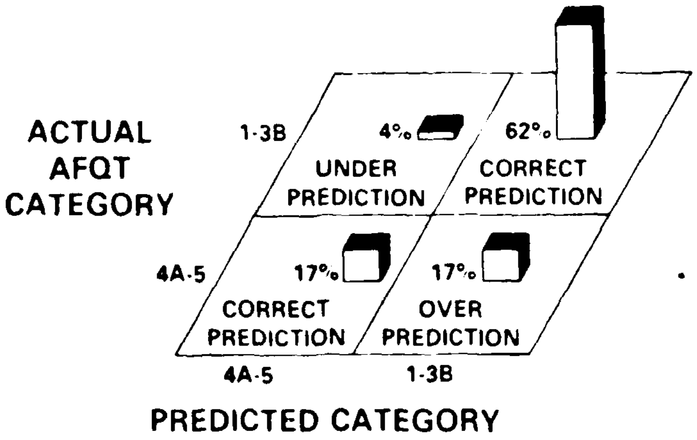

Figure 1a

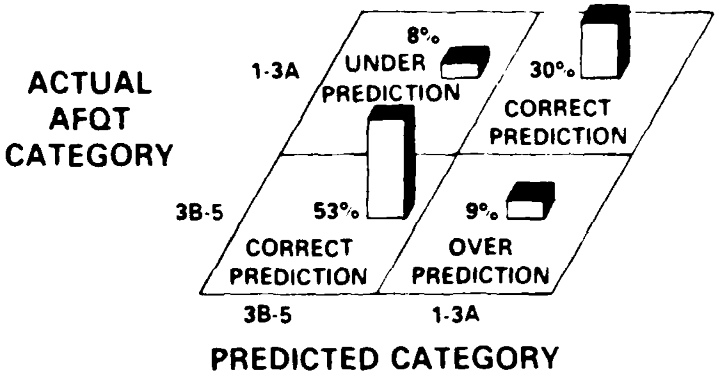

Figure 1b

*NOTE THAT THE PERCENTAGES IN EACH TABLE TOTAL 100

Study National CAST
Cross-validation (1985)

Prepared by
U.S. Army Research Institute

Providing recruiters with probablistic information should improve their ability to use CAST as both an informal screening test and as a sales tool and thereby increase the number of correct AFQT category predictions. Classification analysis could be used to compute the desired probability estimates associated with each CAST score. This would permit CAST to report to the recruiter the estimated probability that the prospect will subsequently be classified into category 3A or higher and the estimated probability that the prospect will subsequently be classified into category 3B or lower. If one probability is much greater than the other (e.g., 85% versus 15%), the recruiter's interpretation of the results is simple. If the two probabilities are close (e.g., 55% versus 45%), then the recruiter can use other considerations, such as distance to MEPS, to decide the appropriate action to take regarding the prospect. These probablities could be presented as shown in Figure 2. Note that the bar charts for the probabilities for the different categories shown in Figure 2 are simply labeled 'A' and 'B' so that recruiters would have some discretion in explaining the meaning of the output to the prospect. 'A' is the probability that the prospect will subsequently be classified into AFQT category 1-3A, and 'B' is the probability that the prospect will subsequently be classified into category 1-3B.

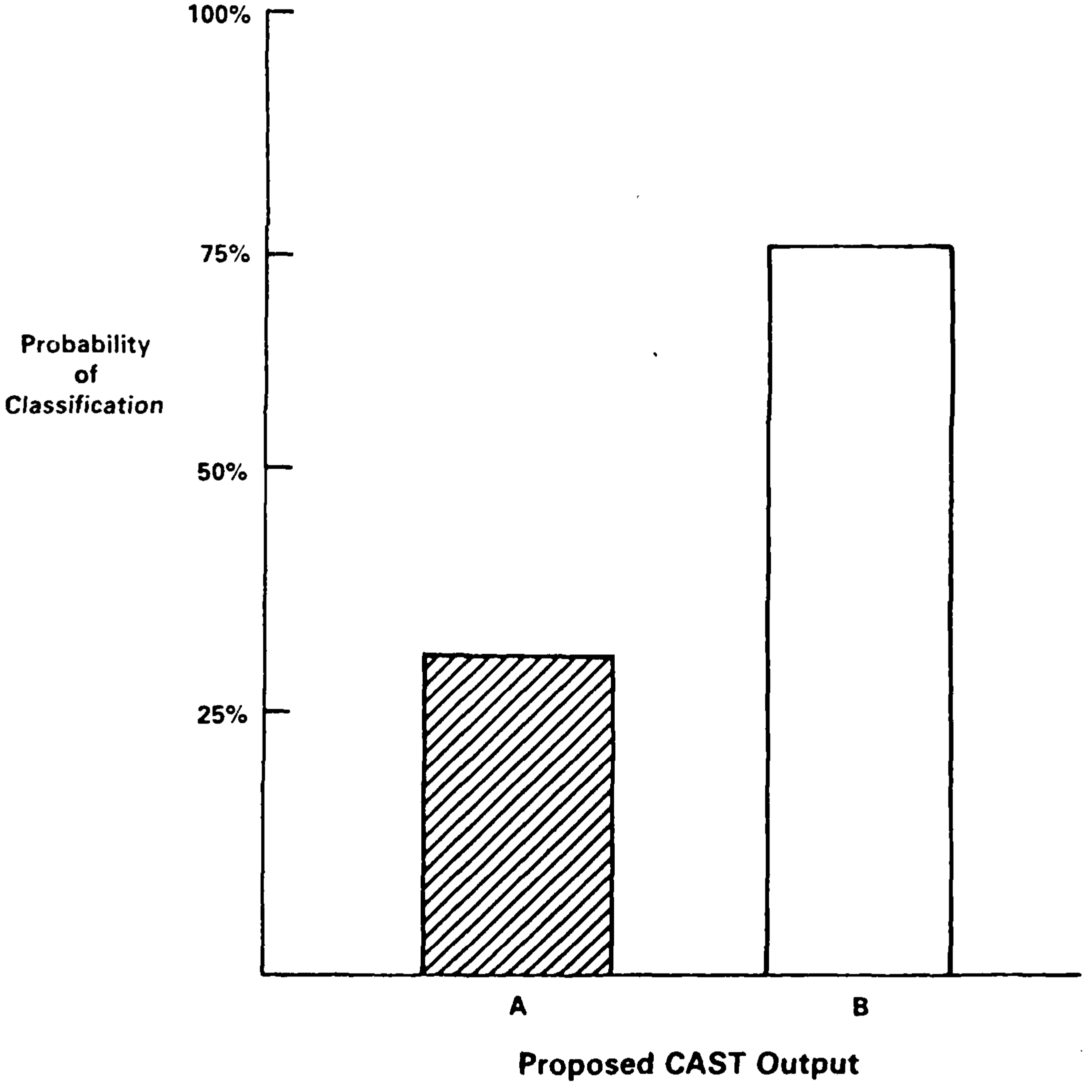
100%
75%
50%
25%
Probability
of
Classification
A
B
Proposed CAST Output

Figure 2

SUMMARY AND RECOMMENDATIONS

In January 1985, 60 Army recruiting stations were asked to begin forwarding CAST and EST data to ARI. This data collection effort continued through December 1985; however, only the data collected during the first six months of 1985 have been analyzed at this time. A final report will be written that will include an anlysis of the entire twelve months of data. Based on the analyses presented in this report, we believe that the current operational version of CAST is quite good at predicting AFQT scores ($\underline{r}^2$ =.67). The efficiency of the current length of the test, 10 WK and 5 AR items, surpasses that of other alternatives (e.g., 10 WK and 10 AR items).

Race and sex subgroup differences in AFQT predictions based on CAST exist, but the magnitude of these differences is not large. The analyses of the CAST data that deal with issues concerning test fairness are summarized in another report (Knapp, et al., 1986). The pattern of differences found in these analyses parallels those found in validity studies of ASVAB (e.g., Dunbar & Novick, 1985; Hanser & Grafton, 1982) and of college entrance examinations like the Scholastic Aptitude Test (e.g., Kallingal, 1971; Temp, 1971).

This research effort also examined the validity of EST, CAST's paper-and-pencil counterpart. Like CAST, EST is highly correlated with AFQT performance ($\underline{r}^2$=.62). EST was also examined for issues relating to test fairness and these analyses are summarized in Knapp et al. (1986). These analyses indicated that there were no significant sex differences in prediction with EST and the racial differences in prediction for EST were similar to those found with CAST.

Currently, recruiters primarily use CAST to predict whether a prospect will be classified into AFQT categories 4A-5, 3B, or 1-3A. Presumably, they accomplish this by interpreting CAST scores exactly as they would interpret AFQT scores. For example, a CAST score of 31 would lead to the prediction that the prospect would be classified into AFQT category 3B. These predictions at the 4A/3B and 3A/3B cutpoints are fairly accurate; however, it is important to note that point predictions of prospects' AFQT scores based on CAST scores are often as much as 14 points above or below their actual AFQT scores. Using classification analysis to provide recruiters with probabilistic information about AFQT category prediction should improve recruiters' ability to make accurate decisions about encouraging the prospect to go on for further testing and about whether or not to sell options and benefits that are only available to "quality" prospects.

Given the current use of CAST as both a sales tool and an informal screening devise, we recommend modifying CAST to provide probability estimates for subsequent AFQT category classification. However, there are two important issues concerning the future use of CAST that must be considered before a final decision about modifying CAST's output is made. First, CAST is being considered for use as a true screening test for all services and probability estimates would not be the preferred type of output for this purpose. Second, changes made to CAST's output will affect future validation efforts. Each of these issues will be discussed in more detail.

Work done within the context of the DoD CAT ASVAB research program has led to the realization that the cost effective operationalization of a CAT ASVAB testing program will require military testing procedures different from those currently used. In particular, widespread use of MET sites for ASVAB testing is not a practical part of a CAT ASVAB testing scenario whereas it is an integral part of current ASVAB testing procedures. One alternative to widespread MET testing that is receiving serious attention is the use by all services of a screening test such as CAST. This test would be used in a "go/no go" fashion so that ASVAB testing would be kept to a minimum. Note that although it would be possible to use CAST in this manner, this step is not justifiable without further research.

The criterion-related validity estimates provided thus far with respect to CAST are useful, but they are not sufficient to justify the use of a specific cutpoint to screen people out of the military service. However, if the additional research was completed, and CAST was used as a true screening test, then a specific cutpoint would be dictated by DoD policy. For example, recruiters would be told to send all prospects who achieve a CAST score of 16 or better on for further testing. In this case, the recruiter would be unable to consider other factors (such as the distance to the nearest MEPS, the recruiter's mission box, etc.) when deciding whether or not to send the prospect on for further testing. The recruiter would not be making a decision, he or she would simply be administering the test and following DoD policy. Therefore, probablistic information about the prospect's subsequent AFQT category classification would not be needed.

The second issue of concern is that modifying CAST to present probabilistic information provided by classification analysis will make it more difficult to validate CAST in the future. This is because an adequate validation effort will require that the recruiters' interpretation of CAST performance be matched to actual AFQT classification to produce information about accuracy rates of actual AFQT category prediction similar to that shown in Figure 1. We are proposing that recruiters be given two probabilities; (1) the probability that the examinee will subsequently be classified into category 3B or above on AFQT, and (2) the probability that the examinee will subsequently be classified into category 3A or above on AFQT. It is then up to the recruiter to decide whether to recommend that the prospect go on for further testing or to discuss options and benefits that are available to "quality" applicants. We recommend that recruiters be instructed to consider other relevant factors, such as the distance to the nearest MEPS and their mission box, when making this decision. In order to assess how accurately recruiters are using CAST to predict AFQT category classification we will have to know what the recruiter decided. Did the recruiter recommend that the prospect go on for further testing? Did the recruiter emphasize to the propsect the existence of incentives that are only available to "quality" applicants.

In light of future research needs, implementation of a detailed and long term CAST and EST record-keeping plan would be advisable. Such a plan should include the assimilation of information regarding prospects' CAST (or EST) performance, race, and sex, as well as action taken by recruiters with respect to tested prospects. This information should be maintained for approximately two years following CAST/EST administration. Records maintained in this fashion would provide data that would be necessary to evaluate CAST for use as a true screening test and/or to evaluate the utility of CAST category probability estimates. In addition, the Uniform Employment Selection Guidelines (1978) published by the Equal Employment Opportunity Commission state that "users of selection procedures . . . should maintain and have available for each job information on adverse impact of the selection process for that job . . . " (p. 38,303). To the extent that CAST and EST are part of the Army's selection system (by virtue of policy and/or practice) records that specify the race and sex of examinees, as well as test performance and selection decisions, should be maintained.

Although the analyses of CAST and EST that examine the performance of important subgroups of the population (e.g., black vs. whites) indicated that no large differences in predictive accuracy exist between subgroups, policy makers need to realize that additional items must be developed for CAST to insure fairness in testing. A valid test is one that accurately measures what it purports to measure. There are a couple of approaches that can be used to examine the validity of a test. The current CAST validation effort, and those preceding it, have used a criterion-related validation paradigm. That is, validity estimates are based on the correlation between CAST (a predictor) and AFQT performance (the criterion). This validation approach is vital to the evaluation of a test such as CAST. Demonstrating that a test predicts what it is supposed to predict is necessary, but not sufficient, for demonstrating that the test measures what it is supposed to measure. This latter question can be answered using a construct validity approach to test development.

Test item development procedures are designed to result in items that are related to the underlying trait or ability that one wishes to measure. CAST attempts to measure two underlying abilities: Word knowledge and arithmetic reasoning. The item calibration procedures used in Item Response Theory (IRT) methodology, the methodology used to construct CAST, are intended to insure that individuals with the same level of the ability being assessed have the same probability of answering a given test item correctly. If these two sets of procedures are carefully conducted, then strong evidence in favor of the resulting test's construct validity will exist. When the test items that currently compose CAST's item pools were calibrated, however, the calibration procedure was performed on all examinees simultaneously. Because items were not also calibrated on examinees grouped by sex and race, there may be items that exhibit construct validity with respect to one subgroup but not another. For example, an item would be racially biased if blacks at a particular ability level are less likely than whites at the same ability level to get the item correct. Although the items in CAST's item pools have not been tested in this fashion, a research effort will begin in the near future to accomplish this task. This effort will result in the elimination of racially-biased items, if there are any, from CAST.

Based on the preliminary findings reported in this paper and in light of the concerns expressed about the future of CAST, the following recommendations are made:

1. Change in CAST's current subtest length is not warranted.

2. Given the current use of CAST by Army recruiters as both a sales tool and an informal screening device, CAST output should be altered so that probabilities associated with a prospect subsequently being classified into one of two critical AFQT categories are reported to the recruiter. These probabilities will be based on the 1980 Youth Norms for AFQT scores.

3. If at some future time, recruiters are given guidelines regarding the interpretation of CAST performance (i.e., a specific "go/no go" cutscore), those guidelines must be based on empirical study. The data presented in the present report are not sufficient to allow for setting such guidelines.

4. Regardless of the intended use of CAST scores at the present time, records containing prospect CAST performance, race, and sex information should be maintained for approximately two years following CAST testing. The conclusions drawn by the recruiter on the basis of CAST performance should also be recorded.

5. All CAST items should be examined for evidence of racial bias via methods based on Item Response Theory. Test items that are found to be unfair to either white or black prospects can then be eliminated from the test.

An ARI research effort beginning in FY86 will address Recommendation 5. The fundamental goal of the new project, however, will be to make any design changes in CAST deemed necessary for the optimization of prediction at critical AFQT cutpoints. To this end, CAST's testing strategy (i.e., ability estimation procedure, item selection rule, and stopping rule) will be reviewed and the WK and AR item pools will be expanded. The expansion of CAST's item pools will also insure their continued integrity.

REFERENCES

Baker, H.G., Rafacz, B.A., & Sands, W.A. (1984). Computerized Adaptive Screening Test (CAST): Development for use in military recruiting stations (NPRDC Report No. 84-17). San Diego, CA: Navy Personnel Research and Development Center.

Birmbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds) Statistical theories of mental test scores. Reading, Mass: Addison-Wesley

Dunbar, S.B. & Novick, M.R. (1985). On predicting success in training for males and females: Marine Corps clerical specialties and ASVAB forms 6 and 7 (ONR Report No. 85-2). Washingtion, DC: Office of Naval Research.

Hanser, L.M. & Grafton, F.C. (1982). Predicting job proficiency in the Army: Race, sex, and education (Selection and Classification WP No.82-1). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Jensema, C.G. (1977). Bayesian tailored testing and the influence of item bank characteristics. Applied Psychological Measurement, 1, 111-120.

Kallingal, A. (1971). The prediction of grades for Black and White students at Michigan State University. Journal of Educational Measurement, 8, 263-266.

Knapp, D.J., Pliske, R.M., & Elig, T.W. (1986). Cross-validation of the Computerized Adaptive Screening Test (CAST): An examination of test fairness (Personnel Utilization Technical Area WP No. 86-11). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Mathews, J.J. & Ree, M.J. (1982). Enlistment Screening Test Forms 81a and 81b: Development and calibration (AFHRL Report No. 81-54). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory.

McBride, J.R. (1979). Adaptive mental testing: The state of the art (ARI Report No. 423). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (NTIS No. 088000)

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Moreno, K.E., Wetzel, C.D., McBride, J.R., & Weiss, D.J. (1983). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests (NPRDC Report No. 83-27). San Diego, CA: Navy Personnel Research and Development Center. (NTIS No. ADA 131683)

Pliske, R.M., Gade, P.A., & Johnson, R.M. (1984). Cross-Validation of the Computerized Adaptive Screening Test (CAST) (ARI Research Report No. 1372). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Sands, W.A., & Gade, P.A. (1983). An application of computerized adaptive testing in U.S. Army Recruiting. Journal of Computer-Based Instruction, 10, 87-89.

Uniform Guidelines on Employee Selection Procedures. Federal Register, 43 (166), 38290-38315 (August 25, 1978).

Temp, G. (1971). Validity of the SAT for Blacks and Whites in thirteen integrated institutions. Journal of Educational Measurement, 8, 245-252.

Warm, T. A. (1978). A primer of item response theory. (USCGI Report No. 941278). Oklahoma City, OK: U.S. Coast Guard Institute.